Loss function landscape

Theories of Deep Learning. Part 3

Eugene Golikov

MIPT, spring 2019

Neural Networks and Deep Learning Lab., MIPT

Spin-glass model

Consider binary classification problem: $y \in \{-1, 1\}$, $x \in \mathbb{R}^{d_0}$;
Hinge loss: $L(y, \hat{y}) = [1 - y\hat{y}]_+$.

Fully-connected feed-forward net:

$$\hat{y} = qW_H\sigma(W_{H-1}\ldots\sigma(W_1 x)\ldots) \in \mathbb{R}^1,$$

or, in case of $\sigma(z) = [z]_+$,

$$\hat{y} = q\sum_{i=1}^{d_0}\sum_{j=1}^{\gamma} x_i a_{i,j}\prod_{k=1}^{H} w_{i,j}^{(k)},$$

where q – some normalizing constant.
Let d_k = width of k-th layer;
$\gamma = \#$paths from input to output $= d_1 \cdot \ldots \cdot d_{H-1}$;
$a_{i,j} \in \{0, 1\}$ – activation of j-th path from i-th input;
$N = d_0 d_1 + \ldots + d_{H-2}d_{H-1} + d_{H-1} \cdot 1$ – number of parameters.

Spin-glass model

Hamiltonian of H-spin spherical spin-glass model (Barrat, 1997[1]):

$$\mathcal{L}_{\Lambda,H}(\mathbf{w}) = \frac{1}{\Lambda^{(H-1)/2}} \sum_{i_1 \ldots i_H = 1}^{\Lambda} J_{i_{1:H}} \prod_{k=1}^{H} w_{i_k} \quad s.t.\ \frac{1}{\Lambda} \sum_{i=1}^{\Lambda} w_i^2 = 1.$$

- Λ – number of spins; H – max. number of interacting spins;
- $\mathbf{w} \in \mathcal{S}^{\Lambda-1}(\sqrt{\Lambda})$ – state of the system;
- $J_{i_{1:H}} \sim \mathcal{N}(0,1)$ – interaction strengths (random variable!).

Features:

- Has multiple critical points;
- Structure of critical points for $\Lambda \to \infty$ is studied in Auffinger et al. (2010)[2] (more about it later).

[1] https://arxiv.org/abs/cond-mat/9701031

[2] https://arxiv.org/abs/1003.1129

Spin-glass model

$$\mathcal{L}_{net}(W) = \mathbb{E}_{x,y\sim\mathcal{D}}\left[1 - yq\sum_{i=1}^{d_0}\sum_{j=1}^{\gamma} x_i a_{i,j}\prod_{k=1}^{H} w_{i,j}^{(k)}\right]_+$$

vs

$$\mathcal{L}_{\Lambda,H}(\mathbf{w}) = \frac{1}{\Lambda^{(H-1)/2}}\sum_{i_1\ldots i_H=1}^{\Lambda} J_{i_{1:H}}\prod_{k=1}^{H} w_{i_k} \quad s.t.\ \frac{1}{\Lambda}\sum_{i=1}^{\Lambda} w_i^2 = 1.$$

Differences:

- deterministic vs random;
- non-negative vs unbounded below;
- N parameters vs Λ parameters.

Similarities:

- (Piece-wise) homogeneous polynomials w.r.t w of degree H.

Spin-glass model

Plan (Choromanska et al., 2014[3]):

1. Introduce a couple of (unrealistic) assumptions to make the loss $\mathcal{L}_{net}(W)$ of the same form as $\mathcal{L}_{\Lambda,H}(\mathbf{w})$;
2. Apply analysis from Auffinger et al. (2010) to reason about "goodness" of critical points.

[3] `https://arxiv.org/abs/1412.0233`

Spin-glass model

$$\hat{y} = q \sum_{i=1}^{d_0} \sum_{j=1}^{\gamma} x_i a_{i,j} \prod_{k=1}^{H} w_{i,j}^{(k)}.$$

Rearrange as:

$$\hat{y} = q \sum_{i=1}^{\Psi} x_i a_i \prod_{k=1}^{H} w_i^{(k)},$$

where $\Psi = d_0 \gamma = \#$paths from all inputs to output.

Assumptions:

- Model hinge loss as Bernoulli RV M:

$$L(y, \hat{y}) = [1 - y\hat{y}]_+ = M(1 - y\hat{y}) = M - q \sum_{i=1}^{\Psi} (yx_i)(Ma_i) \prod_{k=1}^{H} w_i^{(k)}.$$

  Denote $z_i = yx_i$, $b_i = Ma_i$.

Spin-glass model

$$L(y, \hat{y}(x, W)) = M - q \sum_{i=1}^{\Psi} z_i b_i \prod_{k=1}^{H} w_i^{(k)}.$$

Assumptions (Choromanska et al., 2015[4]):

- **A1p:** $b_i \sim \mathrm{Bernoulli}(\rho)\ \forall i$;
- **A2p:** $z_i \sim \mathcal{N}(0, 1)\ \forall i$;
- **A3p:** (weight redundancy) We can leave only $\Lambda = \sqrt[H]{\Psi}$ unique weights $\mathbf{w}$ without sufficient loss of accuracy;
- **A4p:** Every combination of H unique weights appears in the loss.

Given this, we rewrite:

$$L(\mathbf{w}) = M - q \sum_{i_1 \dots i_H = 1}^{\Lambda} z_{i_{1:H}} b_{i_{1:H}} \prod_{k=1}^{H} w_{i_k}.$$

[4] http://proceedings.mlr.press/v40/Choromanska15.pdf

Spin-glass model

$$L(\mathbf{w}) = M - q \sum_{i_1 \ldots i_H = 1}^{\Lambda} z_{i_{1:H}} b_{i_{1:H}} \prod_{k=1}^{H} w_{i_k}.$$

- **A5u:** $b_{i_{1:H}}$ is independent from $z_{i_{1:H}}$:

$$\mathbb{E}_{M, b_{i_{1:H}}} L(\mathbf{w}) = \rho' - q \sum_{i_1 \ldots i_H = 1}^{\Lambda} z_{i_{1:H}} \rho \prod_{k=1}^{H} w_{i_k};$$

- **A6u:** All z_{i_k} are independent;
- **A7p:** Spherical weight constraint: $\sum_{i=1}^{\Lambda} w_i^2 = \Lambda$.

Spin-glass model

$$\mathbb{E}_{M,b_{i_{1:H}}} L(\mathbf{w}) = \rho' - q \sum_{i_1 \dots i_H = 1}^{\Lambda} z_{i_{1:H}} \rho \prod_{k=1}^{H} w_{i_k};$$

Defining $q = -\frac{1}{\rho \Lambda^{(H-1)/2}}$, $J_{i_{1:H}} = z_{i_{1:H}}$, we get:

$$\mathbb{E}_{M,b_{i_{1:H}}} L(\mathbf{w}) = \frac{1}{\Lambda^{(H-1)/2}} \sum_{i_1 \dots i_H = 1}^{\Lambda} J_{i_{1:H}} \prod_{k=1}^{H} w_{i_k} + C \quad s.t. \sum_{i=1}^{\Lambda} w_i^2 = \Lambda.$$

Note: $\Lambda = \sqrt[H]{\Psi} \to \infty \Leftrightarrow N \to \infty$.

Ok, that's spin-glass Hamiltonian . . .

. . . however, we've cheated a lot.

What do we know about energy landscape of spin-glasses?

Spin-glass model

$$\mathcal{L}_{\Lambda,H}(\mathbf{w}) = \frac{1}{\Lambda^{(H-1)/2}} \sum_{i_1 \dots i_H = 1}^{\Lambda} J_{i_{1:H}} \prod_{k=1}^{H} w_{i_k} \quad s.t.\ \frac{1}{\Lambda} \sum_{i=1}^{\Lambda} w_i^2 = 1.$$

Index of critical point $\mathbf{w}$:

$$\operatorname{ind} \mathbf{w} = \#\text{negative eigenvalues of } \nabla^2 \mathcal{L}_{\Lambda,H}(\mathbf{w}).$$

Number of critical points of index k with energy in a set ΛB:

$$\mathcal{C}_{\Lambda,k}(B) = \sum_{\mathbf{w}:\ \nabla \mathcal{L}_{\Lambda,H}(\mathbf{w})=0} [\mathcal{L}_{\Lambda,H}(\mathbf{w}) \in \Lambda B][\operatorname{ind} \mathbf{w} = k], \quad B \subset \mathbb{R};$$

$$\mathcal{C}_{\Lambda,k}(u) := \mathcal{C}_{\Lambda,k}((-\infty, u)).$$

Spin-glass model

$\mathcal{C}_{\Lambda,k}(B)$ is a random variable; Auffinger et al. (2010) derived the following form of its expectation:

$$\mathbb{E}\,\mathcal{C}_{\Lambda,k}(B) = \textit{some complicated expression},$$

which has the following asymptotics for $B = (-\infty, u)$ and $H \geq 2$:

$$\lim_{\Lambda\to\infty} \frac{1}{\Lambda} \log \mathbb{E}\,\mathcal{C}_{\Lambda,k}(u) = \Theta_{k,H}(u),$$

where $\Theta_{k,H}(u)$ is non-decreasing.

Spin-glass model

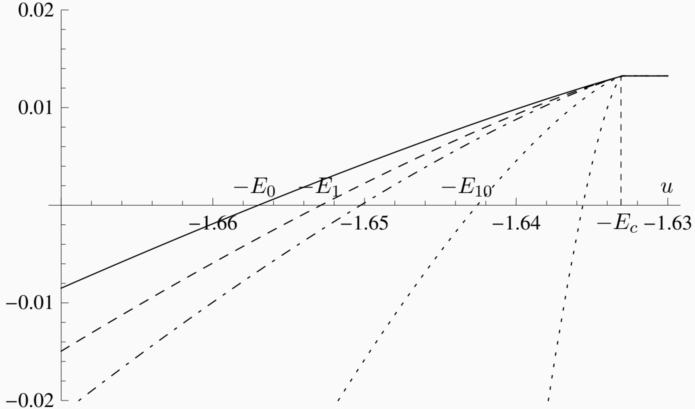

Figure 1: The functions $\Theta_{k,H}$, for $H = 3$ and $k = 0$ (solid), $k = 1$ (dashed), $k = 2$ (dash-dotted), $k = 10$, $k = 100$ (both dotted). All these functions agree for $u \geq -E_\infty$.

Spin-glass model

$$\lim_{\Lambda\to\infty} \frac{1}{\Lambda} \log \mathbb{E}\, \mathcal{C}_{\Lambda,k}(u) = \Theta_{k,H}(u).$$

Define $E_k(H)$ such that $\Theta_{k,H}(-E_k(H)) = 0$.

Properties of $E_k(H)$:

- $\forall H \geq 2\, \forall k \geq 0 \quad E_{k+1}(H) < E_k(H)$;
- $\forall H \geq 2 \quad \lim_{k\to\infty} E_k(H) = E_\infty(H) = 2\sqrt{\frac{H-1}{H}}$.

Spin-glass model

Facts about distribution of critical points of spin-glasses:

In the limit of $\Lambda \to \infty$:

- $-\Lambda E_0(H)$ is an energy of global minimum;
- All critical points of non-diverging index lie within the band $[-\Lambda E_0(H), -\Lambda E_\infty(H)]$;
- All critical points of index k lie within the band $[-\Lambda E_k(H), -\Lambda E_\infty(H)]$;
- Number of local minima in $[-\Lambda E_0(H), -\Lambda E_\infty(H)]$ dominates the number of saddle points in $[-\Lambda E_0(H), -\Lambda E_\infty(H)]$.

Spin-glass model

Idea:
Link the loss surface of a multi-layer network with a Hamiltonian of a physical model.

Why good:

- Connecting ML to physics is cool;
- Tells us why local minima of a loss-surface cannot be arbitrarily bad.

Why bad:

- Applies only to ReLU nonlinearity and Hinge loss;
- Bases on a couple of unrealistic assumptions;
- Length of the band of local minima diverges with Λ (hence with N).

Arbitrary model

Consider finite dataset $(x_i, y_i)_{i=1}^m$, $x_i \in \mathbb{R}^{d_x}$, $y_i \in \mathbb{R}^{d_y}$;
Loss $L(y, \hat{y})$ – convex, differentiable wrt $\hat{y}$.

The most general case:

$$\mathcal{L}(\theta) = \frac{1}{m}\sum_{i=1}^{m} L(y_i, \hat{y}(x_i, \theta)),$$

where $\hat{y}(x, \theta)$ – *any* model.

Introduce modified loss:

$$\tilde{\mathcal{L}}(\theta, W, a, b) = \frac{1}{m}\sum_{i=1}^{m} L(y_i, \hat{y}(x_i, \theta) + a \odot \exp(Wx_i + b)) + \lambda \|a\|_2^2.$$

Arbitrary model

$$\mathcal{L}(\theta) = \frac{1}{m}\sum_{i=1}^{m} L(y_i, \hat{y}(x_i, \theta));$$

$$\tilde{\mathcal{L}}(\theta, W, a, b) = \frac{1}{m}\sum_{i=1}^{m} L(y_i, \hat{y}(x_i, \theta) + a \odot \exp(Wx_i + b)) + \lambda \|a\|_2^2.$$

Theorem (Kawaguchi & Kaelbling, 2019[5]):

If (θ, W, a, b) is a local minimum of $\tilde{\mathcal{L}}$, all $w_{jk}, a_j, b_j \in \mathbb{R}$, then θ is a global minimum of $\mathcal{L}$.

Where is the catch?

[5] `https://arxiv.org/abs/1901.00279`